2005

Theory of mind... for a robot

Brian Scassellati *
MIT Artificial Intelligence Lab
200 Technology Square
Cambridge, MA 02139
scaz@ai.mit.edu
http://www.ai.mit.edu/people/scaz/

Abstract

One of the fundamental social skills for humans is a theory of other minds. This set of skills allows us to attribute beliefs, goals, and desires to other individuals. To take part in normal human social dynamics, a robot must not only know about the properties of objects, but also the properties of animate agents in the world. This paper presents the theories of Leslie (1994) and Baron-Cohen (1995) on the development of theory of mind in human children and discusses the potential application of both of these theories to building robots with similar capabilities. Initial implementation details and basic skills (such as finding faces and eyes and distinguishing animate from inanimate stimuli) are introduced. We further speculate on the usefulness of a robotic implementation in evaluating and comparing these two models.

Introduction

Human social dynamics rely upon the ability to correctly attribute beliefs, goals, and percepts to other people. This set of metarepresentational abilities, which have been collectively called a "theory of mind", allows us to understand the actions and expressions of others within an intentional or goal-directed framework (what Dennett (1987) has called the intentional stance). The recognition that other individuals have knowledge, perceptions, and intentions that differ from our own is a critical step in a child's development and is believed to be instrumental in self-recognition, grounding in linguistic acquisition, and possibly in the development of imaginative and creative play (Byrne & Whiten 1988). These abilities are also central to what defines human interactions. Normal social interactions depend upon the recognition of other points of view, the understanding of other mental states, and the recognition of complex non-verbal signals of attention and emotional state.

Research from many different disciplines have focused on theory of mind. Students of philosophy have been interested in the understanding of other minds and the representation of knowledge in others for the past two centuries. Most recently, Dennett (1987) has focused on how organisms naturally adopt an "intentional stance" and interpret the behaviors of others as if they possess goals, intents, and beliefs. Ethologists have also focused on the issues of theory of mind. Studies of the social skills present in primates and other mammals have revolved around the extent to which other species are able to interpret the behavior of conspecifics and influence that behavior through deception (e.g. Premack (1988), Povinelli and Preuss (1995), and Cheney and Seyfarth (1991)). Research on the development of social skills in children have focused on characterizing the developmental progression of social abilities (e.g. Fodor (1992), Wimmer and Perner (1983), and Frith and Frith (1999)) and on how these skills result in conceptual changes and the representational capacities of infants (e.g. Carey (1999) and Gelman (1990)). Furthermore, research on pervasive developmental disorders such as autism have focused on the selective impairment of these social skills (e.g. Perner and Lang (1999), Karmiloff-Smith et. al. (1995), and Mundy and Sigman (1989)).

Researchers studying the development of social skills in normal children, the presence of social skills in primates and other vertebrates, and certain pervasive developmental disorders have all focused on attempting to decompose the idea of a central "theory of mind" into sets of precursor skills and developmental modules. In this abstract, I will attempt to review two of the most popular and influential general models which attempt to link together multi-disciplinary research into a coherent developmental explanation, one from Baron-Cohen (1995) and one from Leslie (1994). I will then describe the initial phases of a research program aimed at implementing many of these precursor skills on a humanoid robot.

Leslie's model of theory of mind

Leslie's (1984) theory treats the representation of causal events as a central organizing principle to theories of object mechanics and theories of other minds much in the same way that the notion of number may be central to object representation. According to Leslie, the world is naturally decomposed into three classes of stimuli based upon their causal structure; one class for *mechanical agency*, one for *actional agency*, and one for *attitudinal agency*. Leslie argues that evolution has produced independent domain-

*Parts of this research are funded by DARPA/ITO under contract number DABT 63-99-1-0012 and parts have been funded by ONR under contract number N00014-95-1-0600, "A Trainable Modular Vision System."

specific modules to deal with each of these classes of event. The Theory of Body module (ToBY) deals with events that are best described by mechanical agency, that is, they can be explained by the rules of *mechanics*. The second module is system 1 of the Theory of Mind module (ToMM-1) which explains events in terms of the intent and goals of agents, that is, their *actions*. The third module is system 2 of the Theory of Mind module (ToMM-2) which explains events in terms of the *attitudes* and beliefs of agents.

The first mechanism is the Theory of Body mechanism (ToBY) which embodies the infant's understanding of physical objects. ToBY is a domain-specific module that deals with the understanding of physical causality in a mechanical sense. ToBY's goal is to describe the world in terms of the mechanics of physical objects and the events they enter into. ToBY is believe to operate on two types of visual input: a three-dimensional object-centered representation from high level cognitive and visual systems and a simpler motion-based system. This second system accounts for the causal explanations that adults give (and the causal expectations of children) to the "billiard ball" type launching displays pioneered by Michotte (1962). Leslie proposed that this mechanism is innate, but more recent work from Cohen and Amsel (1998) may show that it develops extremely rapidly in the first few months and is fully developed by 6.5 months.

ToBy is followed developmentally by the emergence of a Theory of Mind Mechanism (ToMM) which develops in two phases, which Leslie calls system-1 and system-2 but which we will refer to as ToMM-1 and ToMM-2 after Baron-Cohen (1995). Just as ToBY deals with the physical laws that govern objects, ToMM deals with the psychological laws that govern agents. ToMM-1 is concerned with actional agency; it deals with agents and the goal-directed actions that they produce. This system of detecting goals and actions begins to emerge at around 6 months of age, and is most often characterized by attention to eye gaze. Leslie leaves open the issue of whether ToMM-1 is innate or acquired. ToMM-2 is concerned with attitudinal agency; it deals with the representations of beliefs and how mental states can drive behavior relative to a goal. This system develops gradually, with the first signs of development beginning between 18 and 24 months of age and completing sometime near 48 months. ToMM-2 employs the M-representation, a metarepresentation which allows the truth properties of a statement to be based on mental states rather than observable stimuli. ToMM-2 is a required system for understanding that others hold beliefs that differ from our own knowledge or from the observable world, for understanding different perceptual perspectives, and for understanding pretense and pretending.

Baron-Cohen's model of theory of mind

Baron-Cohen's model assumes two forms of perceptual information are available as input. The first percept describes all stimuli in the visual, auditory, and tactile perceptual spheres that have self-propelled motion. The second percept describes all visual stimuli that have eye-like shapes. Baron-Cohen proposes that the set of precursors to a theory of mind, which he calls the "mindreading system," can be decomposed into four distinct modules.

The first module interprets self-propelled motion of stimuli in terms of the primitive volitional mental states of goal and desire. This module, called the intentionality detector (ID) produces dyadic representations that describe the basic movements of approach and avoidance. For example, ID can produce representations such as "he wants the food" or "she wants to go over there". This module only operates on stimuli that have self-propelled motion, and thus pass a criteria for distinguishing stimuli that are potentially animate (agents) from those that are not (objects). Baron-Cohen speculates that ID is a part of the innate endowment that infants are born with.

The second module processes visual stimuli that are eye-like to determine the direction of gaze. This module, called the eye direction detector (EDD), has three basic functions. First, it detects the presence of eye-like stimuli in the visual field. Human infants have a preference to look at human faces, and spend more time gazing at the eyes than at other parts of the face. Second, EDD computes whether the eyes are looking at it or at something else. Baron-Cohen proposes that having someone else make eye contact is a natural psychological releaser that produces pleasure in infants (but may produce more negative arousal in other animals). Third, EDD interprets gaze direction as a perceptual state, that is, EDD codes dyadic representational states of the form "agent sees me" and "agent looking-at not-me."

The third module, the shared attention mechanism (SAM), takes the dyadic representations from ID and EDD and produces triadic representations of the form "John sees (I see the girl)." Embedded within this representation is a specification that the external agent and the self are both attending to the same perceptual object or event. This shared attentional state results from an embedding of one dyadic representation within another. SAM additionally can make the output of ID available to EDD, allowing the interpretation of eye direction as a goal state. By allowing the agent to interpret the gaze of others as intentions, SAM provides a mechanism for creating nested representations of the form "John sees (I want the toy)."

The last module, the theory of mind mechanism (ToMM), provides a way of representing epistemic mental states in other agents and a mechanism for tying together our knowledge of mental states into a coherent whole as a usable theory. ToMM first allows the construction of representations of the form "John believes (it is raining)." ToMM allows the suspension of the normal truth relations of propositions (referential opacity), which provides a means for representing knowledge states that are neither necessarily true nor match the knowledge of the organism, such as "John thinks (Elvis is alive)." Baron-Cohen proposes that the triadic representations of SAM are converted through experience into the M-representations of ToMM.

For normal children, Baron-Cohen proposes that both ID and the basic functions of EDD are available to infants in the first 9 months of life. SAM develops between 9 and 18 months, and ToMM develops from 18 months to 48 months. One of the most attractive parts of this model though is the ways in which it has been applied both to the abnormal development of social skills in autism and the ways that the de-

composition has been compared to the social skills of other primates and vertebrates.

Autism is a pervasive developmental disorder of unknown etiology that is diagnosed by a checklist of behavioral criteria. Baron-Cohen has proposed that the range of deficiencies in autism can be characterized by his model. In all cases, EDD and ID are present. In some cases of autism, SAM and ToMM are impaired, while in others only ToMM is impaired. This can be contrasted with other developmental disorders (such as Down's syndrome) or specific linguistic disorder in which evidence of all four modules can be seen.

Furthermore, Baron-Cohen attempts to provide an evolutionary description of these modules by identifying partial abilities in other primates and vertebrates. This phylogenetic description ranges from the abilities of hog-nosed snakes to detect direct eye contact to the sensitivities of chimpanzees to intentional acts. Roughly speaking, the abilities of EDD seem to be the most basic and can be found in part in snakes, avians, and most other vertebrates as a sensitivity to predators (or prey) looking at the animal. ID seems to be present in many primates, but the capabilities of SAM seem to be present only partially in the great apes. The evidence on ToMM is less clear, but it appears that no other primates readily infer mental states of belief and knowledge.

Implementing a theory of mind for a robot

A robotic system that possessed a theory of mind would allow for social interactions between the robot and humans that have previously not been possible. The robot would be capable of learning from an observer using normal social signals in the same way that human infants learn; no specialized training of the observer would be necessary. The robot would also be capable of expressing its internal state (emotions, desires, goals, etc.) through social interactions without relying upon an artificial vocabulary. Further, a robot that can recognize the goals and desires of others will allow for systems that can more accurately react to the emotional, attentional, and cognitive states of the observer, can learn to anticipate the reactions of the observer, and can modify its own behavior accordingly. The construction of these systems may also provide a new tool for investigating the predictive power and validity of the models from natural systems that serve as the basis. An implemented model can be tested in ways that are not possible to test on humans, using alternate developmental conditions, alternate experiences, and alternate educational and intervention approaches.

The difficulty, of course, is that even the initial components of these models require the coordination of a large number of perceptual, sensory-motor, attentional, and cognitive processes. In this section, I will outline the advantages and disadvantages of Leslie's model and Baron-Cohen's model with respect to implementation. In the following section, I will describe some of the components that have already been constructed and some which are currently designed but still being implemented.

From a robotics standpoint, the most salient differences between the two models are in the ways in which they divide perceptual tasks. Leslie cleanly divides the perceptual world into animate and inanimate spheres, and allows for further processing to occur specifically on each type of stimulus. Baron-Cohen does not divide the perceptual world quite so cleanly, but does provide more detail on limiting the specific perceptual inputs that each module requires. In practice, both models require remarkably similar perceptual systems (which is not surprising, since the behavioral data is not under debate). However, each perspective is useful in its own way in building a robotic implementation. At one level, the robot must distinguish between object stimuli that are to be interpreted according to physical laws and agent stimuli that are to be interpreted according to psychological laws. However, the specifications that Baron-Cohen provides will be necessary for building visual routines that have limited scope.

The implementation of the higher-level scope of each of these models also has implications to robotics. Leslie's model has a very elegant decomposition into three distinct areas of influence, but the interactions between these levels are not well specified. Connections between modules in Baron-Cohen's model are better specified, but they are still less than ideal for a robotics implementation. Issues on how stimuli are to be divided between the competencies of different modules must be resolved for both models. On the positive side, the representations that are constructed by components in both models are well specified.

Components of a Robotic Theory of Mind

Taking both Baron-Cohen's model and Leslie's model, we can begin to specify the perceptual and cognitive abilities that our robots must employ. Our initial systems concentrate on two abilities: distinguishing between animate and inanimate motion and on identifying gaze direction. To maintain engineering constraints, we must focus on systems that can be performed with limited computational resources, at interactive rates in real time, and on noisy and incomplete data. To maintain biological plausibility, we focus on building systems that match the available data on infant perceptual abilities.

We have constructed an upper-torso humanoid robot with a pair of six degree-of-freedom arms, a three degree-of-freedom torso, and a seven degree of freedom head and neck. The robot, named Cog, has a visual system consisting of four color CCD cameras (two cameras per eye, one with a wide field of view and one with a narrow field of view at higher acuity), an auditory system consisting of two microphones, a vestibular system consisting of a three axis inertial package, and an assortment of kinesthetic sensing from potentiometers, strain gauges, and thermistors.

We are currently implementing a system that distinguishes between animate and inanimate visual stimuli based on the presence of self-generated motion. Similar to the findings of Leslie (1982) and Cohen and Amsel (1998) on the classification performed by infants, our system operates at two developmental stages. Both stages form trajectories from stimuli in consecutive image frames and attempt to maximize the path coherency. This computational technique for multi-target tracking has been used extensively in signal processing domains, and our approach is most similar to the algorithm proposed by Reid (1979) and implemented

by Cox and Hingorani (1996). We are currently developing metrics for evaluating these trajectories in order to classify the stimulus as either animate or inanimate using the descriptions of Michotte's observations of adults and Leslie's observations of infants. The differences between the two developmental states lies in the type of features used in tracking. At the first stage, representing infants before 6 months of age, only spatio-temporal features (resulting from object size and motion) are used as cues for tracking. In the second stage, more complex object features such as color, texture, and shape are employed. With a system for distinguishing animate from inanimate stimuli, we can begin to provide the distinctions implicit in Leslie's differences between ToBY and ToMM and the assumptions that Baron-Cohen requires for ID.

The first shared attention behaviors that infants engage in involve maintaining eye contact. To enable our robot to recognize and maintain eye contact, we have implemented a perceptual system capable of finding faces and eyes (Scassellati 1998). The system first locates potential face locations using a template-based matching algorithm developed by Sinha (1996). Once a potential face location has been identified, the robot saccades to that target using a learned visual-motor behavior. The location of the face in peripheral image coordinates is then mapped into foveal image coordinates using a second learned mapping. The location of the face within the peripheral image can then be used to extract the sub-image containing the eye for further processing. This technique has been successful at locating and extracting sub-images that contain eyes under a variety of conditions and from many different individuals. These functions match the first function of Baron-Cohen's EDD and begin to approach the second and third functions as well. We are currently extending the functionality to include interpolation of gaze direction using the decomposition proposed by Butterworth (1991).

In addition to these obvious behaviors, there are also a variety of behavioral and cognitive skills that are not integral parts of the theory of mind models, but are nonetheless necessary to implement the desired functionality. We have implemented a variety of perceptual feature detectors (such as color saliency detectors, motion detectors, skin color filters, and rough disparity detectors) that match the perceptual abilities of young infants. We have constructed a model of human visual search and attention that was proposed by Wolfe (1994). We have also implemented motor control schemes for visual motor behaviors (including saccades, smooth-pursuit tracking, and a vestibular-occular reflex), orientation movements of the head and neck, and primitive reaching movements for a six degree-of-freedom arm.

At this workshop, I will present the implementation of these basic social skills and discuss the usefulness of models of theory of mind from Baron-Cohen and Leslie in implementing robotic systems.

References

Baron-Cohen, S. 1995. *Mindblindness.* MIT Press.

Butterworth, G. 1991. The ontogeny and phylogeny of joint visual attention. In Whiten, A., ed., *Natural Theories of Mind.* Blackwell.

Byrne, R., and Whiten, A., eds. 1988. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford University Press.

Carey, S. 1999. Sources of conceptual change. In Scholnick, E. K.; Nelson, K.; Gelman, S. A.; and Miller, P. H., eds., *Conceptual Development: Piaget's Legacy.* Lawrence Erlbaum Associates. 293–326.

Cheney, D. L., and Seyfarth, R. M. 1991. Reading minds or reading behavior? Tests for a theory of mind in monkeys. In Whiten, A., ed., *Natural Theories of Mind.* Blackwell.

Cohen, L. B., and Amsel, G. 1998. Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Develoment* 21(4):713–732.

Cox, I. J., and Hingorani, S. L. 1996. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 18(2):138–150.

Dennett, D. C. 1987. *The Intentional Stance.* MIT Press.

Fodor, J. 1992. A theory of the child's theory of mind. *Cognition* 44:283–296.

Frith, C. D., and Frith, U. 1999. Interacting minds – a biological basis. *Science* 286:1692–1695.

Gelman, R. 1990. First principles organize attention to and learning about relevant data: number and the animate-inanimate distinction as examples. *Cognitive Science* 14:79–106.

Karmiloff-Smith, A.; Klima, E.; Bellugi, U.; Grant, J.; and Baron-Cohen, S. 1995. Is there a social module? Language, face processing, and theory of mind in individuals with Williams Syndrome. *Journal of Cognitive Neuroscience* 7:2:196–208.

Leslie, A. M. 1982. The perception of causality in infants. *Perception* 11:173–186.

Leslie, A. M. 1984. Spatiotemporal continuity and the perception of causality in infants. *Perception* 13:287–305.

Leslie, A. M. 1994. ToMM, ToBY, and Agency: Core architecture and domain specificity. In Hirschfeld, L. A., and Gelman, S. A., eds., *Mapping the Mind: Domain specificity in cognition and culture.* Cambridge University Press. 119–148.

Michotte, A. 1962. *The perception of causality.* Andover, MA: Methuen.

Mundy, P., and Sigman, M. 1989. The theoretical implications of joint attention deficits in autism. *Development and Psychopathology* 1:173–183.

Perner, J., and Lang, B. 1999. Development of theory of mind and executive control. *Trends in Cognitive Sciences* 3(9).

Povinelli, D. J., and Preuss, T. M. 1995. Theory of mind: evolutionary history of a cognitive specialization. *Trends in Neuroscience* 18(9).

Premack, D. 1988. "Does the chimpanzee have a theory of mind?" revisited. In Byrne, R., and Whiten, A., eds., *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans.* Oxford University Press.

Reid, D. B. 1979. An algorithm for tracking multiple targets. *IEEE Transactions on Automated Control* AC-24(6):843–854.

Scassellati, B. 1998. Finding eyes and faces with a foveated vision system. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98).*

Sinha, P. 1996. *Perceiving and recognizing three-dimensional forms.* Ph.D. Dissertation, Massachusetts Institute of Technology.

Wimmer, H., and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–128.

Wolfe, J. M. 1994. Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review* 1(2):202–238.